Computational Analysis of Leukemia Microarray Expression Data Using the GA/KNN Method

Leping Li,[1,*] Lee G. Pedersen,[2] Thomas A. Darden,[3] and Clarice R. Weinberg[1]

[1] Biostatistics Branch and [3]Laboratory of Structural Biology, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, USA
[2]Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599-3290, USA

*Correspondence should be addressed to L.L. (voice: 919-541-5168, fax: 919-541-7880, email: Li3@niehs.nih.gov)

Running title: sample classification using the GA/KNN.
Keywords: leukemia, pattern recognition, genetic algorithm, k-nearest neighbors, microarray.

Abstract

We recently developed a multivariate method that selects a subset of discriminative genes for sample classification based on gene expression data. The method combines a search tool, a genetic algorithm (GA), and a non-parametric pattern recognition method, based on the k-nearest nearest neighbors (KNN). We begin by selecting many subsets of genes that can discriminate among classes of samples using a training set. Subsequently, the genes are ranked according to the frequency of gene selection. The top-ranked genes (e.g. 50) are then used to classify test set samples. For a widely-available set of leukemia data, the top 50 genes identified by the GA/KNN method not only correctly classified 33 of the 34 test set samples, but also discovered the two distinct clinical subtypes within ALL without applying prior knowledge. The method has been successfully applied to several expression data sets. It may be used to identify a subset of informative genes (biomarkers) for sample classification for a variety of profiling studies including tumors.

Introduction

Microarray technology has made it possible to monitor the global gene expression patterns of cells. It has been used to profile the gene expression of normal and transformed human cells in several tumors including colon (Alon *et al.*, 1999), leukemia (Golub *et al.*1999), prostate (Bubendorf *et al.*, 1999), breast (Perou *et al.*, 2000), lymphoma (Alizadeh *et al.*, 2000), and melanoma (Bittner *et al.*, 2000) as well as the NCI's 60 human tumor cell lines (Ross *et al.*, 2000). These studies may shed light on the mechanisms of cell maltransformation. Therefore, they may be useful in elucidating the mechanisms involved in carcinogenesis and in identifying characteristic patterns for cancer diagnosis and classification.

Methods that mine large expression data sets for sample classification have been reported (Golub et al., 1999, Ben-Dor et al., 2000, Bittner et al., 2000, Li et al., 2001a & 2001b, Perou et al., 2000). The GA/KNN method that we developed is a multi-dimensional classification method. It not only selects a small subset of genes that *jointly* discriminates among classes of samples (e.g. normal vs. tumor) but also assesses the relative predictive importance of all genes for sample classification (Li et al., 2001a & 2001b). We begin by selecting a subset of genes that can discriminate between different classes of samples. When many such subsets of differentially expressed genes are obtained independently using the GA/KNN method, the relative importance of genes for sample classification can be assessed by examining the frequency of selection of the genes into these near-optimal subsets. The most informative set of genes (class predictor) is subsequently used to classify an "unknown" (validation) set of samples. We have applied the method to several data sets including a colon data set (Alon et al., 1999), a lymphoma data (Alizadeh et al., 2000), and as well as a toxicogenomics data set generated at the National Institute of Environmental Health Sciences (to be published). Herein we report the detailed analysis of the leukemia data (Golub et al., 1999) using the GA/KNN method to find sets of genes that can distinguish between acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML).

Methods

Data set. The original leukemia data were downloaded from the web (http://www.genome.wi.mit.edu/MPR). The data contain the expression levels of 6,817 genes across 72 samples, of which 47 had been classified as ALL and 25 as AML (Golub et al., 1999). We excluded any genes with expression levels below 50 in more than 80% of the 72 samples. The number of genes remaining after this filtration is 5545. The expression values were then transformed as follows.

$$f(x) = \begin{cases} c + \log_{10}(x) & x > 1 \\ c & x \geq -1 \text{ and } x \leq 1 \\ -c - \log_{10}(\text{abs}(x)) & x < -1 \end{cases}$$

where *x* is the expression level, *c* is a constant and was set to 0.

We divided the data set into a training set (first 38 samples) and a test set (34 samples) following Golub et al. (1999). The training set was used to obtain a subset of genes that can discriminate between AML and ALL. The 50 most informative genes obtained using the training set were subsequently used in validation, to predict the classifications of the test set samples.

The GA/KNN method. Details of the GA/KNN method have been reported (Li et al., 2001a) and are available on the web (http://chun.niehs.nih.gov/~leping). The method employs a non-parametric pattern recognition method, the k-nearest neighbor (KNN) (see e.g., Vandeginste et al., 1998), and a searching tool, a genetic algorithm (GA) (for a review, see Judson 1997). The KNN method was used as the classification tool that distinguishes between discriminative and non-discriminative sets of genes while the GA was used to choose a relatively few subsets of genes (from many possible combinations) for testing.

In brief, each sample is represented by a pattern of expression that consists of *d* genes. Each sample is then KNN classified according to the class memberships of its *k* (arbitrarily set to 3) nearest neighbors, as determined by the Euclidean distance in the *d*-dimensional space. If all of the 3 nearest neighbors of a sample are ALL, the sample is classified as ALL; similarly for AML. If the 3 nearest neighbors are not of the same class, the sample is considered unclassifiable.

The GA is used to search high-dimensional space, since selecting an optimal subset of optimal genes from a large gene pool is a combinatorial problem. Initially, a population of randomly selected "chromosomes" is generated. Each "chromosome" consists of *d* genes. The "fitness" (merit) of each "chromosome" is determined by its ability to classify the training set samples. If the class memberships of a training set sample and its three nearest neighbors in the particular *d*-dimensional space defined by a "chromosome" agree, a score of 1 is assigned to that sample. These agreement scores are summed across the training set and we refer to this sum divided by the number of samples (here 38) as the fitness score. Thus, the GA is used to optimize the fitness score.

A set of *d* genes was considered discriminative in which each sample and its three nearest neighbors are of the same class for at least 37 of the 38 training set samples. To study the sensitivity of gene selection to the choices of *d*, we systematically examined various sizes of *d* (e.g., 5, 10, 20, 30, 40, 50, 70, and 100). For each *d*, 20,000 discriminative sets of genes (referred to as the near-optimal "chromosomes") that potentially discriminate between AML and ALL were obtained. Subsequently, the genes were ranked according to the frequency of selection into near-optimal "chromosomes". The most frequently selected genes (e.g. top 50) were then used to classify the test set samples.

Results

Frequency of gene selection and choice of *d*

The statistical *z* score, based on normalizing the frequency with which each of the 5,455 genes was selected in the 20,000 solutions for each *d*, is shown in **Figure 1**. A few genes dominate the selection when *d* is small (e.g. 5). As *d* increases, the pattern of gene selection stabilizes. The correlation between the *z* scores for two choices of *d* is shown in **Figure 2**.

To examine the reproducibility of gene selection, an additional 20,000 near-optimal "chromosomes" were obtained independently for each *d*. Again, each near-optimal "chromosome" contains *d* genes that can discriminate between AML and ALL for the training set samples. The correlation between the *z* scores for two independent runs for each *d* is shown in **Figure 3**. The reproducibility is apparently high for all *d* except *d* = 100. For large *d* (e.g. 100), gene selection is likely contaminated with noise. Furthermore, the calculation becomes computationally expensive as *d* increases. Taken together, a choice of *d* between 20 and 70 appears to yield a stable gene selection for the leukemia data set. We chose a *d* of 40, taking reassurance from the fact that the selection of optimal genes is evidently insensitive to this choice.

Top genes

The genes were ranked based on the frequency of gene selection (*d* = 40) and the top 50 genes are listed in **Table 1**. It is noteworthy that many more were selected with high z scores. The complete list of 5,455 genes based on frequency rank is available on the web (http://chun.niehs.nih.gov/~leping).

Among the top 50 genes, several were myeloid, T-lymphocytic- or B-lymphocyte-specific antigens or surface markers. For instance, CCAAT/enhancer binding protein δ (C/EBPδ) is a transcriptional regulator involved in myeloid cell differentiation (Sterneck et al., 1998). Myeloperoxidase (MPO) is a well-established marker of myeloid differentiation (Austin et al., 1998; Winterbourn 2000). The CD2 is a T-cell antigen that has been used as a diagnosis marker for T-cell ALL (Manabe et al., 1998). It plays an important role in mediating the interactions between human T lymphocytes and accessory cells (Brown et al., 1989). The B-cell lymphocyte kinase, Blk, is a src-family protein tyrosine kinase that is expressed preferentially in B lineage cells (Malek et al., 1998). Blk may play an important role in B cell proliferation (Malek et al., 1998). Immunoglobulin-associated beta (B29), also called CD79b, belongs to a family of surface adhesion molecules on B lymphocytes. B29 and the associated transmembrane protein mb1 are crucial for assembly and membrane display of the B-cell receptor (Hombach et al., 1990). Mutation of *B29* has been associated with chronic lympocytic leukemia (Gordon et al., 2000). The octamer binding factor 1 (OBF-1) is a B-lymphocyte-specific transcription factor (Schubart et al., 1996; Matthias 1998). It has been suggested that over-expression of *OBF*-1 might play a role in the pathogenesis of germinal center-derived B cell lymphoma (Greiner et al., 2000). CD19 is also B lymphoid antigen and used as a diagnosis marker for B-cell ALL (Scheuermann & Racila, 1995). The *TCL1* is an oncogene located on chromosome 14 band q32.1 that has been implicated in the development of mature T cell leukemia (Virgilio et al., 1993; Pekarsky et al., 2000).

Several oncogenes or tumor antigens were also frequently selected, for instance, the proviral integration oncogene *spi1*, *GRO2* oncogene, and carcinoembryonic antigen precursor gene.

Classification of the test set samples

The GA/KNN method correctly classified 33 of the 34 test samples using three training set neighbors (2 or 3 must agree) by KNN using the top 50 genes (**Table 2**). When the 100 top-ranked genes were used, a similar result was obtained except that sample AML54 was classified as ALL. As more genes were included, contaminating the system with high-dimensional noise, the number of misclassified samples increased. In fact, when all 5,455 genes were used, 9 of the 34 test set samples were incorrectly classified. This result emphasizes that not all expression data are relevant to the distinction between ALL and AML. Interestingly, nearly all the misclassified samples were AML, suggesting that the ALL samples are relatively easy to classify. Among those that were misclassified, AML66 was consistently classified as ALL (**Table 2**). When all 72 samples were used for training, AML66 remained classified as ALL (data not shown). Together, these results show that AML66 is an outlier.

When the test set was combined with the training set, a postprocessing cluster analysis of the top 50 genes using a cluster analysis program (Eisen et al., 1998) showed that the AML samples and ALL samples were clustered together correctly except AML66 (**Figure 4**). Furthermore, the top 50 genes found by the GA/KNN method revealed the existence of two subtypes within ALL without applying any prior knowledge. Among the 47 ALL samples, 9 has been clinically classified as T-cell ALL (ALL2, ALL3, ALL6, ALL9, ALL10, ALL11, ALL14, ALL23, and ALL67), and the remaining as B-cell ALL. When clustered using the top 50 genes, each of the 9 T-cell ALL samples were on one branch of the ALL tree together with two B-cell ALL. These results indicate that the GA/KNN method is capable of identifying genes that discriminate not only between the ALL and AML, but that the method may also unmask clinically meaningful subtypes, through subsequent cluster analysis.

We also applied principal component (PC) analysis (see, e.g., Vandeginste et al., 1998) to the 72 samples using the top 50 genes. The method of single value decomposition [(SVD) from Numerical Recipes (Cambridge, MA)] was applied to the correlation matrix. The principal components were obtained by projecting the original data points that have been transformed (see methods) and then mean centered onto the eigenvectors. When plotted the 72 samples using the first two principal components, two distinct clusters are apparent with AML66 in the wrong cluster (**Figure 5**).

In addition, we studied whether each of the 40,000 individual near-optimal "chromosomes" could classify as well as the top 50 genes. Each test set sample was classified 40,000 times as either AML or ALL using 3 nearest *training* set neighbors by the KNN method using a majority rule (2 or 3 must agree). Interestingly, three AML samples (AML54, AML60, and AML66) were classified as ALL more than 50% of the time (**Figure 6**). Among the three, AML66 was misclassified 93% of the time. Furthermore, none of the 40,000 individual near-optimal "chromosomes" correctly classified more than 32 of the 34 test set samples. This result suggests that the individual near-optimal "chromosomes" are not as effective as the set of top-ranked genes (e.g. the top 50) that were selected from these "chromosomes". This instability may be due to inherent noise present in these "chromosomes".

Discussion

It is clear that not all genes are relevant to sample discrimination. Thus, identification of a subset of informative genes is important. The current approaches for selecting a subset of informative genes can be summarized by two categories, univariate and multivariate. An univariate approach examines one gene at a time whereas a multivariate approach considers several genes simultaneously. The widely used discriminative measures include the Student's *t* statistics or similar measures (Golub et al., 1999), discriminant analysis (see, e.g. Vandeginste *et al*., 1998), neural network (Toronen et al., 1999), support vector machines (for a tutorial, see Cortes & Vapnik, 1995; Ben-Dor et al., 2000;), and nearest neighbors (Li et al, 2001a). Alternatively, when selection of individual genes is not carried out, a few "supergenes" are constructed as the informative genes using multivariate approaches such as the principal component analysis (PCA).

Univariate approaches may fail to identify sets of genes that are discriminative *jointly*, but not singly. We have shown that the frequency of co-selection of certain sets of genes (e.g. P_{ABC} for genes A, B, and C) can be many times higher than the product of their individual frequency of selection ($P_A \times P_B \times P_C$) (Li et al., 2001a & 2001b), indicating the multivariate nature of their contribution to discrimination.

Many supervised pattern recognition methods have been employed as discrimination tools. Some methods (such as discriminant analysis) define a boundary between classes explicitly while others (such as the KNN) do so implicitly. This distinction may have an implication for sample classification, especially when samples can be heterogeneous in nature, such as tumors. It is clear that tumor subtypes may exist within known classes. Discriminative functions that can work well in the presence of such heterogeneity are preferred. Subsequently, hidden subcategories in the sample may be discovered and may prove to be etiologically or prognostically important.

The last important issue is that microarray data consist of a large number of genes (parameters) and a small number of samples and as a result, many distinct and equally effective classifiers may exist for the same training set. Most of current literature methods seek a single subset of discriminative genes. Often, the informative genes identified for a given data set vary from method to method.

The GA/KNN method selects many subsets of discriminative genes from which a single predictor set is formed by examining the frequency of gene selection. This simple approach of gene selection appears to function well. The top-ranked genes (e.g. the top 50) apparently outperform the individual subsets of genes. Furthermore, the GA/KNN method accommodates the heterogeneity within the classes, since the KNN method finds a boundary between classes *implicitly*. This unique ability of KNN facilitates subclass discovery.

In conclusion, many methods have been developed for sample classification based on gene expression data. As the quantitative aspects of the microarray technology improve and computational methods that mine the resulting large data set are developed further, the technology will have a great impact on biology, toxicology and related areas.

Acknowledgements

Part of the work was completed when LL was employed at the National Institute for Occupational Safety and Health at Morgantown, West Virginia. LL would like to thank David Umbach for advice on data transformation and insightful discussions and Shyamal Peddada for critical reading the manuscript, and the North Carolina Supercomputing Center for access to their facilities. Most of the computations were carried out on SGI workstations.

References

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. *Proc. Natl. Acad. Sci. USA,***1999**, *96*, 6745.

Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X., Powell,J.I., Yang,L., Marti,G.E., Moore,T., Hudson,J.Jr, Lu,L., Lewis,D.B., Tibshirani,R., Sherlock,G., Chan,W.C., Greiner,T.C., Weisenburger,D.D., Armitage,J.O., Warnke,R., Levy,R., Wilson,E., Grever,M.R., Byrd,J.C., Botstein,D., Brown,P.O. and Staudt, L.M. *Nature*, **2000**, *403*, 503.

Austin,G.E., Alvarado,C.S., Austin,E.D., Hakami,N., Zhao,W.G., Chauvenet,A., Borowitz,M.J., and Carroll,A.J. *Am. J. Clin. Pathol*., **1998**, *110*, 575.

Ben-Dor,A., Bruhn,L., Friedman,N., Nachman,I., Schummer,M., Yakhini,Z. and Ben-Dor,A. In *Proceedings of the Fourth International Conference on Computational Molecular Biology (RECOMB2000*), ACM press, New York, **2000**.

Bittner,M., Meitzer,P., Chen,Y., Jiang,Y., Seftor,E., Hendrix,M., Radmacher,M., Simon,R., Yakhini,Z., Ben-Dor,A., Sampas,N., Dougherty,E., Wang,E., Marincola,F., Gooden,C., Lueders,J., Glatfelter,A., Pollock,P., Carpten,J., Gillanders,E., Leja,D., Dietrich,K., Beaudry,C., Berens,M., Alberts,D., Sondak,V., Hayward,N. and Trent,J. *Nature*, **2000**, *406*, 536.

Brown,M.H., Cantrell,D.A., Brattsand,G., Crumpton,M.J., and Gullberg,M. *Nature*, **1989**, *339*, 551.

Cortes,C. and Vapnik,V. *Machine Learning*, **1995**, *20*, 273.

Eisen,M.B., Spellman,P.T., Brown,P.O., and Botstein, D. *Proc. Natl. Acad. Sci. USA,* **1998**, *95*, 14863.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H, Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D., and Lander,E.S. *Science,* **1999***, 286*, 531.

Gordon,M.S., Kato,R.M., Lansigan,F., Thompson,A.A., Wall,R., and Rawlings,D.J. *Proc. Natl. Acad. Sci. USA*, **2000**, *97*, 5504.

Greiner,A., Muller,K.B., Hess,J., Pfeffer,K., Muller-Hermelink,H.K., and Wirth,T. *Am. J. Pathol*., **2000**, *56*, 501.

Hombach,J., Lottspeich,F., and Reth,M. *Eur. J. Immunol*., **1990**, *20*, 2795.

Hombach,J., Tsubata,T., Leclercq,L., Stappert,H. and Reth,M. *Nature*, **1990**, *343*, 760.

Judson,R. Genetic algorthms and their use in chemistry. *In* Lipkowitz,K.B. and Boyd,D.B. (eds), *Reviews in Computational Chemistry,* VCH publishers, New York, **1997,** vol 10, pp 1-66,

Li,L., Darden,T.A., Weinberg,C.R. and Pedersen,L.G. *Comb. Chem. High Throughput Screen*., accepted, **2001a**.

Li,L, Pedersen, L.G., Darden, T.A. and Weinberg, C.R. *Bioinformatics*, *submitted*, **2001b**

Malek,S.N., Dordai,D.I., Reim,J., Dintzis,H., and Desiderio,S. *Proc. Natl. Sci. Acad. USA*, **1998**, *95*, 7351.

Manabe,A., Mori,T., Ebihara,Y., Koyama,T., Okuyama,I., Hosoya,R., Kaneko,M., Ishimoto,K., Nakahata,T., and Nakazawa,S. *Inter. J. Hematol*., **1998**, *67*, 45.

Massart,D.L., Vandeginste,B.G.M., Deming,S.N., Michotte,Y., and Kaufman, L. In *Chemometrics*: *a textbook (Data Handling in Science and Technology, vol 2)*; Elsevier Science B. V: New York, **1988**, pp. 339-368.

Matthias P. *Semin. Immunol.*, **1998**, *10*, 155.

Pekarsky,Y., Koval,A., Hallas,C., Bichi,R., Tresini,M., Malstrom,S., Russo,G., Tsichlis,P., and Croce,C.M. *Proc. Natl. Acad. Sci. USA*, **2000**, *97*, 3028.

Perou,C.M., Sørlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Aksien,L.A., Fluge,O., Pergamenschikov,A., Williams,C., Zhu,S.X., Lonning,P.E., Børresen-Dale,A.L., Brown,P.O. and Botstein,D. *Nature*, **2000**, *406*, 747.

Schubart,D.B., Rolink,A., Kosco-Vilbois,M.H., Botteri,F., and Matthias,P. *Nature*, **1996**, *383*, 538.

Scheuermann,R.H. and Racila,E. *Leuk. Lymphoma*, **1995**, *18*, 385.

Sterneck,E., Paylor,R., Jackson-Lewis,V., Libbey,M., Przedborski,S., Tessarollo,L., Crawley,J.N., and Johnson,P.F. Proc. *Natl. Sci. Acad. USA*, **1998**, *95*, 10908.

Toronen.P, Kolehmainen.M, Wong.C, and Castren.E. *FEBS lett.*, **1999**, *451*, 142.

Winterbourn,C.C., Vissers,M.C.M., and Kettle,A.J. *Curr. Opin. Hematol.*, **2000**, *7*, 53.

Vandeginste,B.G.M., Massart,D.L., Buydens,L.M.C., De Jong,S., Lewi,P.J. and Smeyers-Verbeke,J. In *Handbook of Chemometrics and Qualimetrics, Part B,* Elsevier Science, The Netherlands, **1998**.

Virgilio,L, Isobe,M, Narducci,MG, Carotenuto,P, Camerini,B, Kurosawa,N, Abbas-ar-Rushdi, Croce,CM, and Russo,G. *Proc. Natl. Acad. Sci. USA*, **1993**, *90*, 9275.

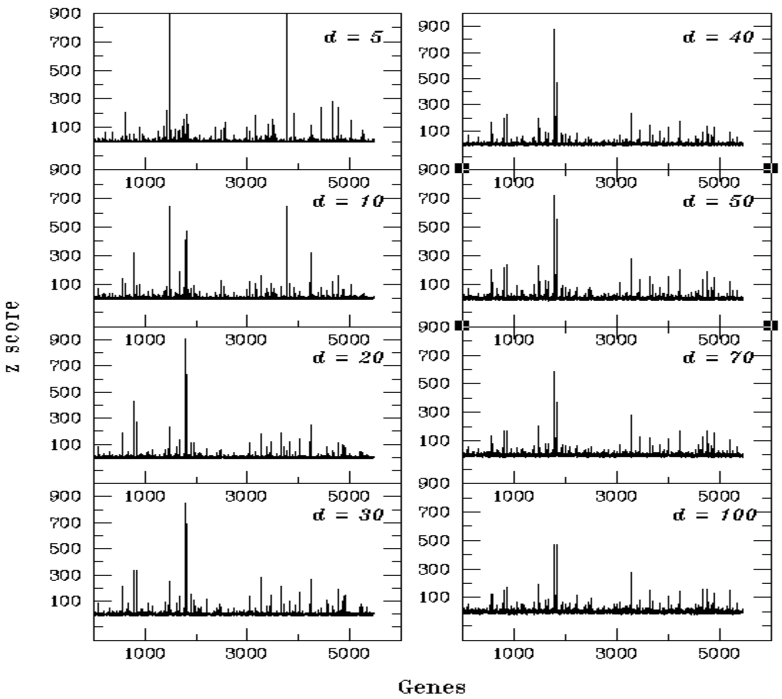

Fig. 1 The statistical *z*-score with which each of the 5,455 genes is selected among the 40,000 solutions. Let, $Z = [S_i - E(S_i)]/\sigma$, where S_i is the number of times gene$_i$ is selected, $E(S_i)$, is the expected number of times gene$_i$ is selected, σ is the square root of the variance. Let, A = 40,000, P(gene$_i$) = *d*/5455, the probability of gene$_i$ being selected (if random). Then, $E(S_i) = P(\text{gene}_i)\cdot A$, and $\sigma = \sqrt{\{P(\text{gene}_i)\cdot[1-P(\text{gene}_i)]\cdot A\}}$. The two highest peaks in *d* = 5 were truncated.

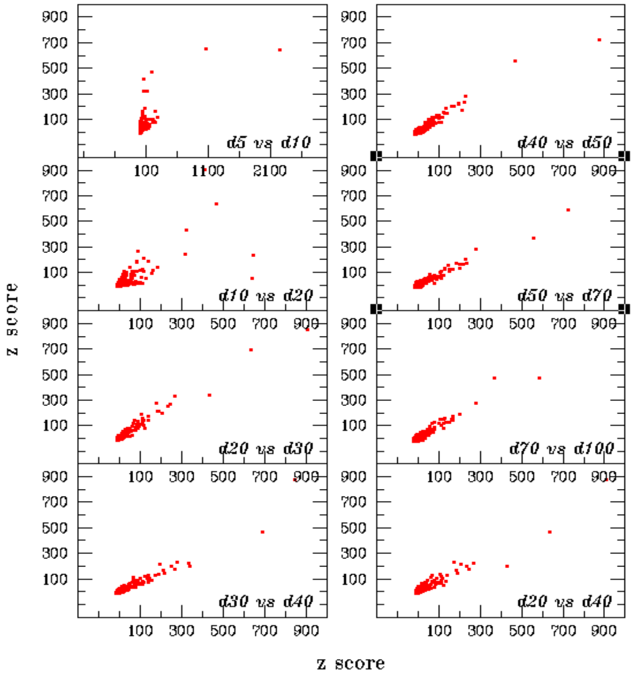

Fig. 2. Correlation between the *z* scores for two consecutive choices of *d*. The z scores were calculated as described in Fig. 1.

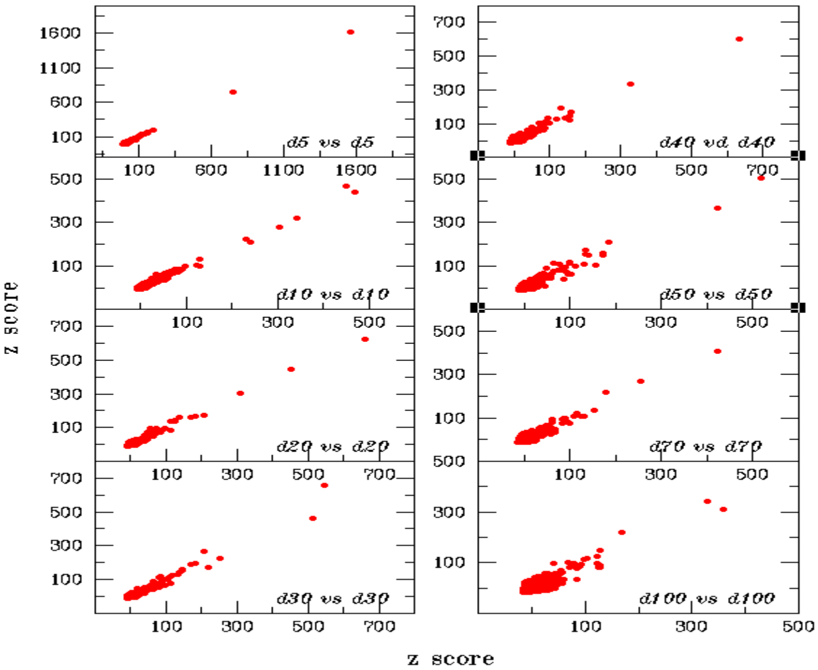

Fig. 3. Reproducibility of gene selection as represented by the correlation between the *z* scores from two independent runs for the same *d*. For each run, 20,000 solutions were obtained for each *d*.

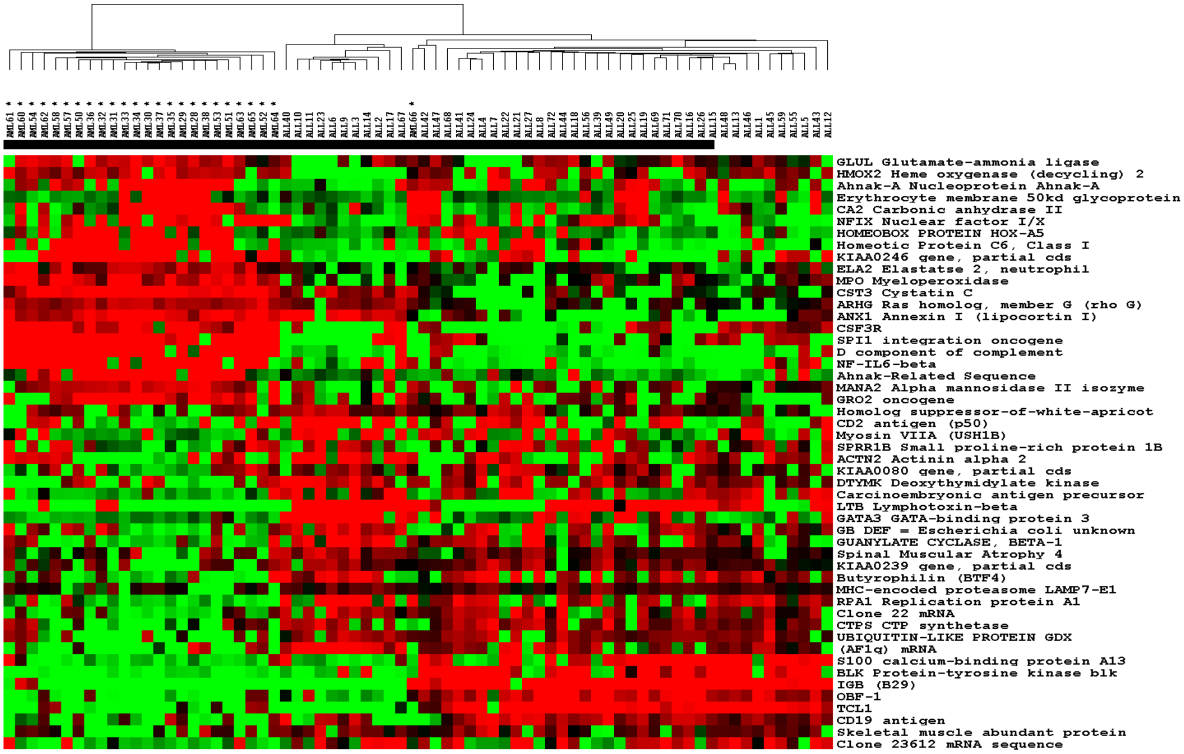

Fig. 4. Clustering analysis (Eisen et al., 1998) of the 72 leukemia samples and the 50 top-ranked genes that were selected from all 5,455 genes by the GA/KNN method.

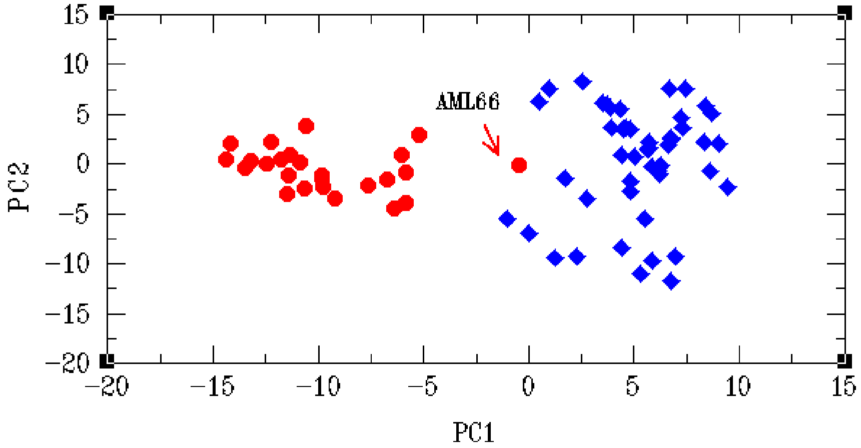

Fig. 5. Principal component analysis of the 72 samples based on the 50 top-ranked genes. The first two principal components are shown.

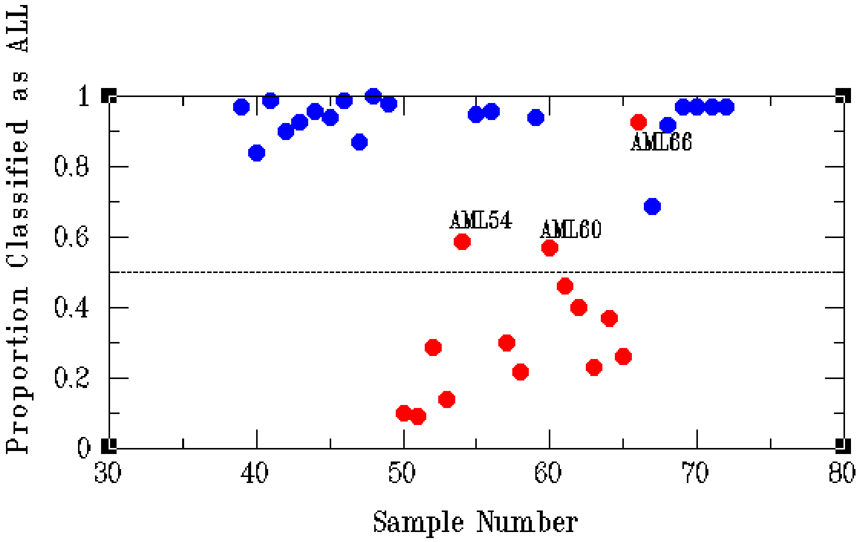

Fig. 6. Percentage of being classified as ALL for the test set samples. Each test set sample was classified by three *training* set neighbors using the top 50 genes by KNN method. A majority rule (2 or 3 must agree) was applied. Each sample was classified 40,000 times using each of the 40,000 solutions (near-optimal "chromosomes") obtained during the training process (see text for details).